SAPPHIRE-based clustering

Supporting information

Francesco Cocina, Andreas Vitalis,* and Amedeo Caflisch

Biochemistry department, University of Zurich, Zurich, Switzerland

E-mail: a.vitalis@bioc.uzh.ch

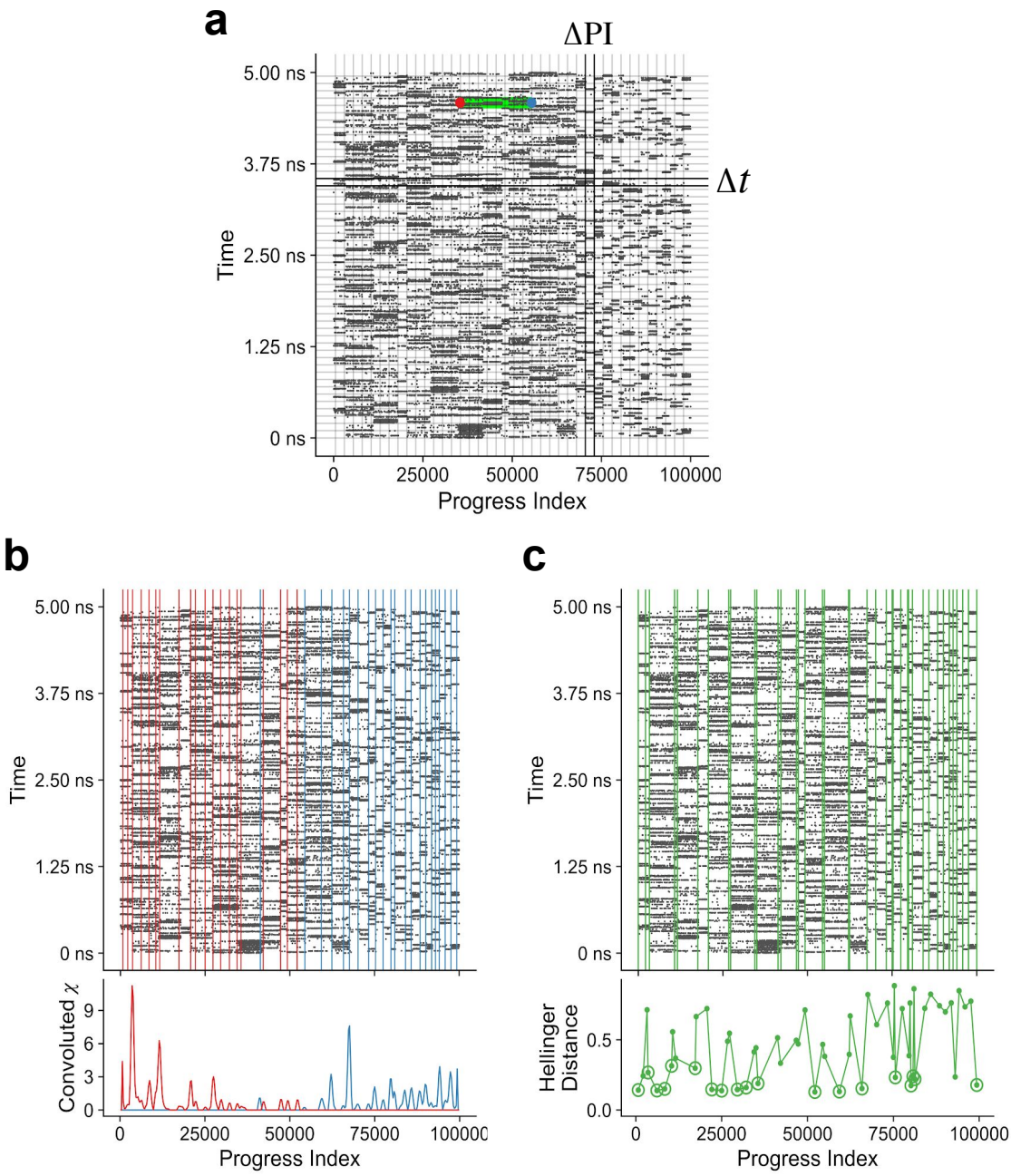

Figure S1. Clustering from the temporal annotation as described in Sec. 2.2.1 for n-butane. (a) A 2-D histogram is created with bin sizes ΔPI and Δt for the PI (x-axis) and 'Time' (y-axis) values, respectively. Stretches of similar occupancy along the PI axis are identified along each row of the histogram (the green segment serves as a single example). Note that the coarsness of the binning on the time axis influences the ability to isolate relatively short visits of individual states. The minima and maxima of the stretches (for the highlighted example, these are the red and blue dots) are then used to compute two different functions as follows. (*continued*)

Figure S1. (*continued*) (b) First, the frequencies of the occurrence of end points of segments along the PI axes are smoothed by an integration/differentiation technique (lower panel). The red and blue distributions correspond to left and right end points, respectively. The peaks in these two smoothed profiles are used to derive an initial partitioning (blue and red vertical lines in the upper panel). Note that the identity of the green stretch shown in (a) is no longer directly relevant for the partitioning in (b), *i.e.* only its boundaries matter, but they do so independently. (c) Second, Hellinger distances are computed between the temporal distributions of each adjacent pair of putative partitions (lower panel). The resulting value is compared to a null model distribution obtained by numerical simulation. A Grubbs test is used to determine if the actual difference value is a significant outlier, *i.e.*, if the distributions are indeed sufficiently dissimilar to infer that they indicate visits of different states. Differences deemed insignificant are shown as circled dots in the lower panel, and only the retained partitions are shown as vertical green lines in the upper panel.

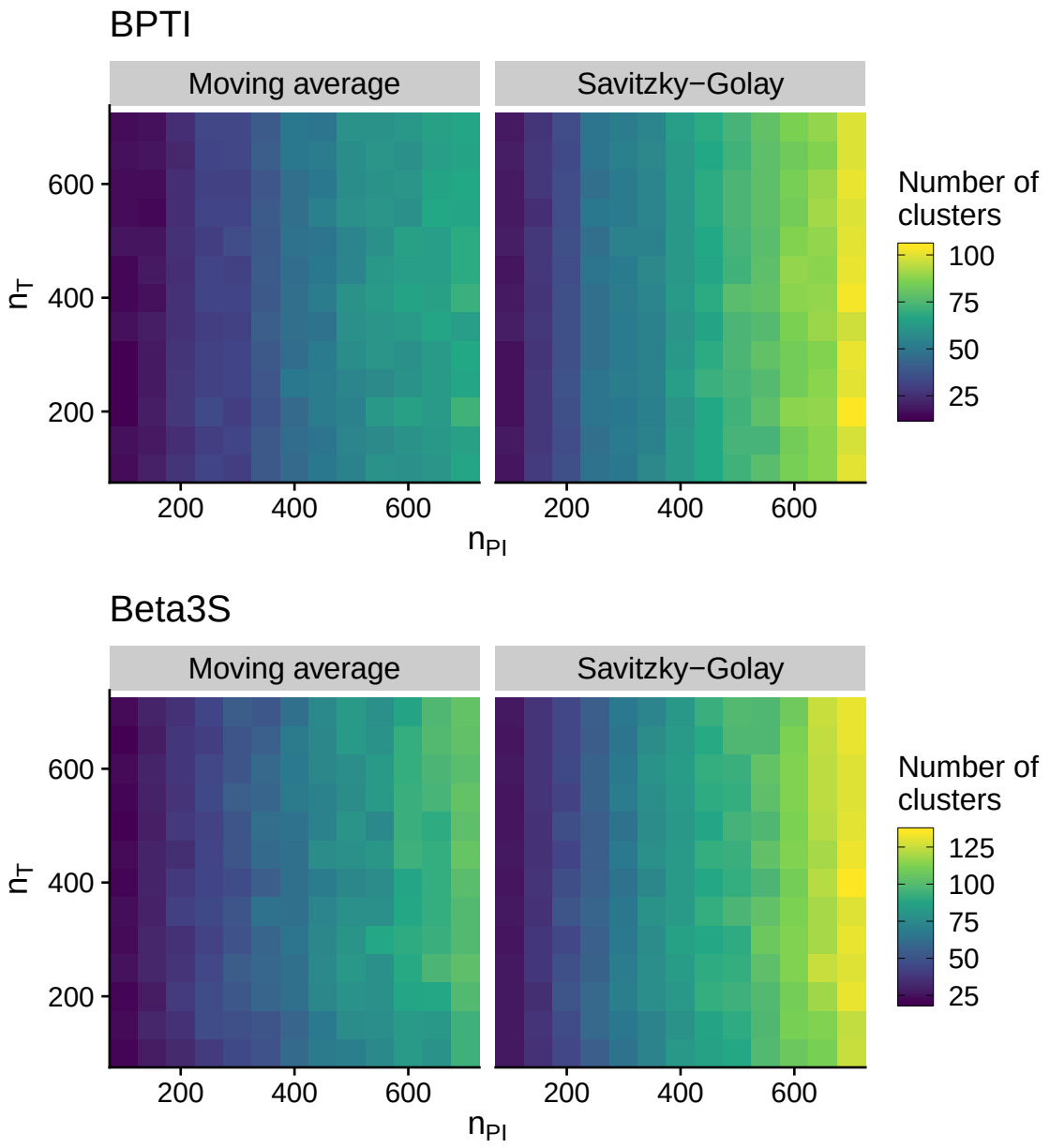

Figure S2. Number of states obtained by SAPPHIRE-based clustering (SbC) for different settings. We show here how the number of clusters varies with respect to the number of bins along the PI axis, n_{PI}, to the number of bins along the 'Time' axis, n_T, and to the type of smoothing filter adopted for the kinetic annotation analysis. The filter was either a moving average one or a Savitzky-Golay filter with degree two.

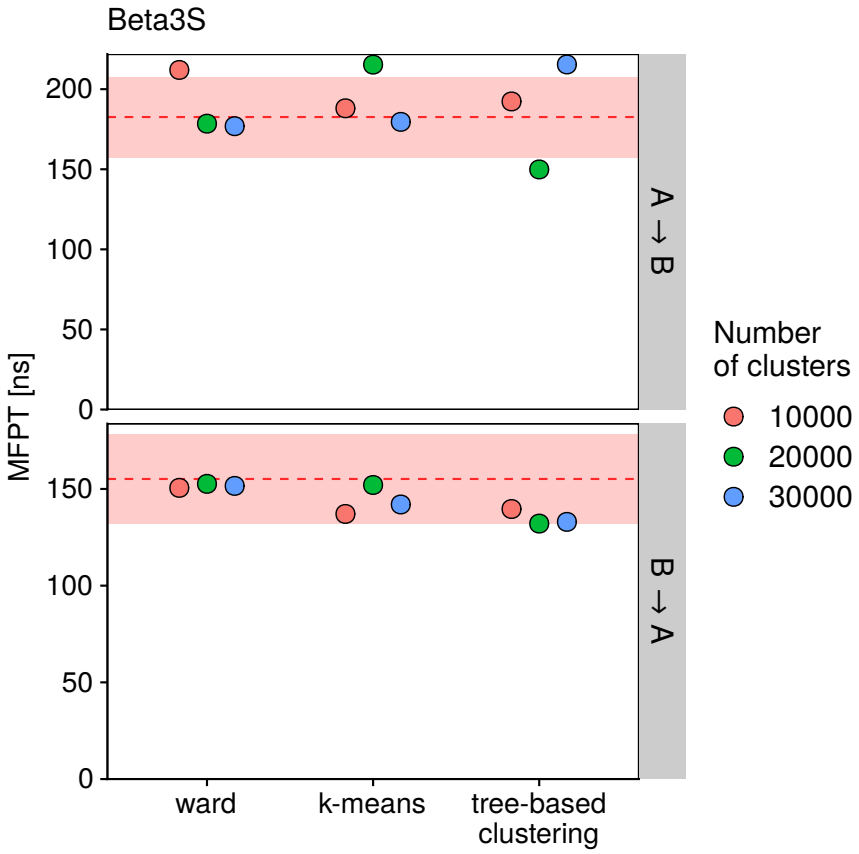

Figure S3. Mean-first passage times (MFPTs) for Beta3S for larger numbers of clusters. The circles show results for the 'Direct' MFPT estimates computed on the discrete trajectory and can be compared directly to the red box plots in Fig. 6 in the main text. The dotted lines and shaded regions represent MFPT estimates on the original sets of labels and their standard errors, respectively. These latter data are the same as in Fig. 6. 'Tree-based clustering' is a highly efficient, tree-based clustering method suitable for large numbers of clusters.[1] The construction of the short spanning tree (sST) actually relies on this clustering technique as mentioned in Sec. 2.1 of the main text. The data demonstrate that discretization errors are primarily to blame for the lack of agreement between the various direct estimates in Fig. 6.

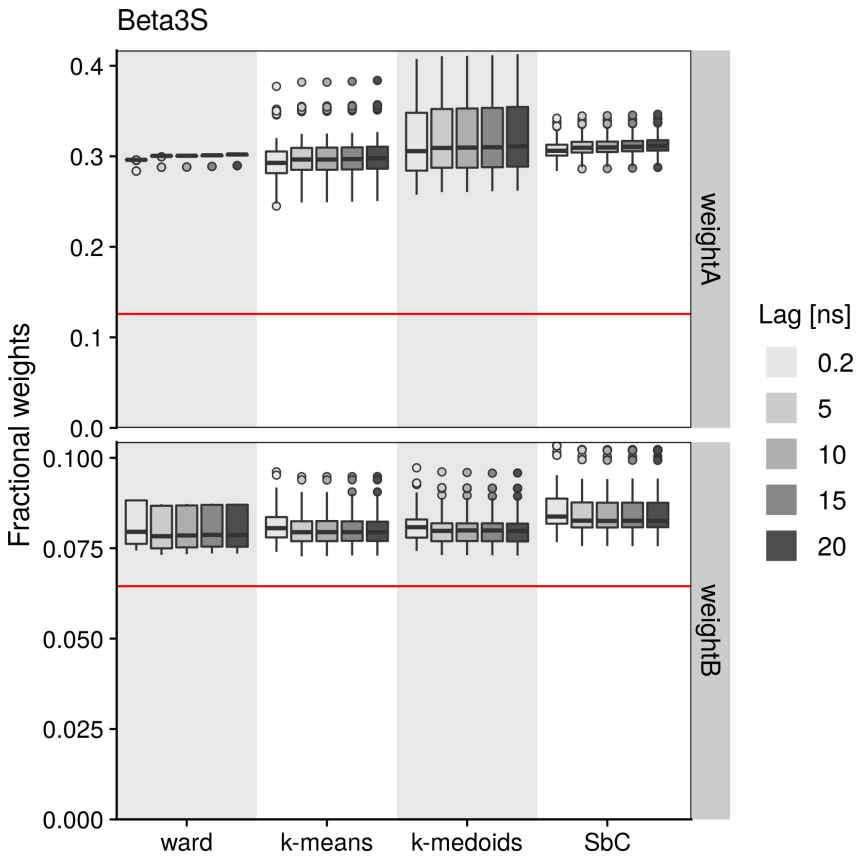

Figure S4. Statistical weight of macro-states A and B for Beta3S. Box plots indicate the total weight of the macro-states inferred from the stationary distribution of the underlying MSM. The models shown here are the same as in Fig. 6b. Red lines indicate the effective weight of the macro-states extracted from the original trajectory based on labels alone. The differences arise because the DSSP-based criterion is so strict that many of the snapshots within large and relatively homogeneous clusters are not labeled. This is particularly true for macro-state A.

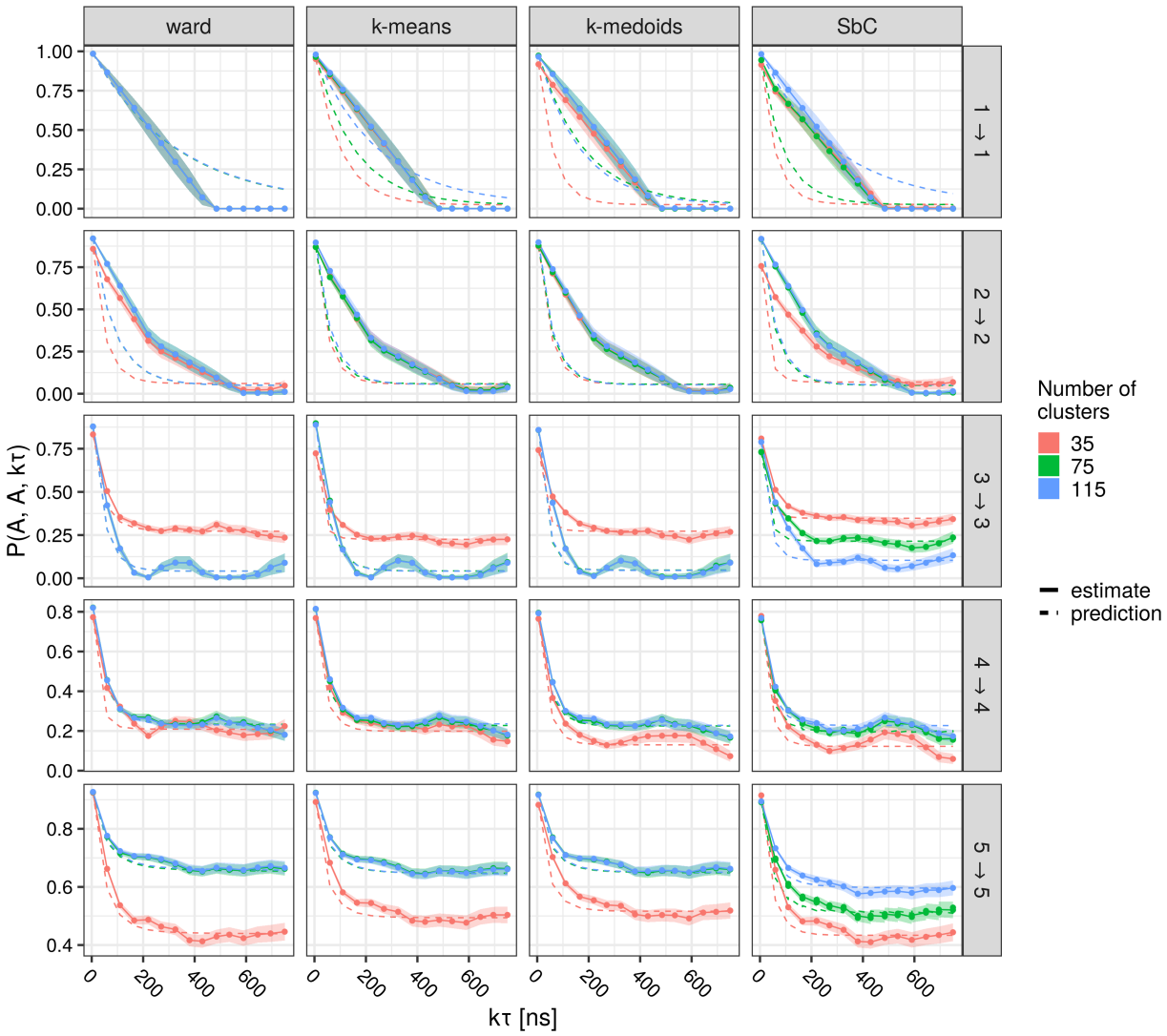

Figure S5. Chapman-Kolmogorov test of Beta3S for different clustering methods. Coarse-graining of the trajectories into five macrostates was performed with PCCA+.[2] Predictions are computed from an unconstrained maximum likelihood MSM with lag equal to 5 ns. Standard errors for the estimated values are shown as shaded regions.[3]

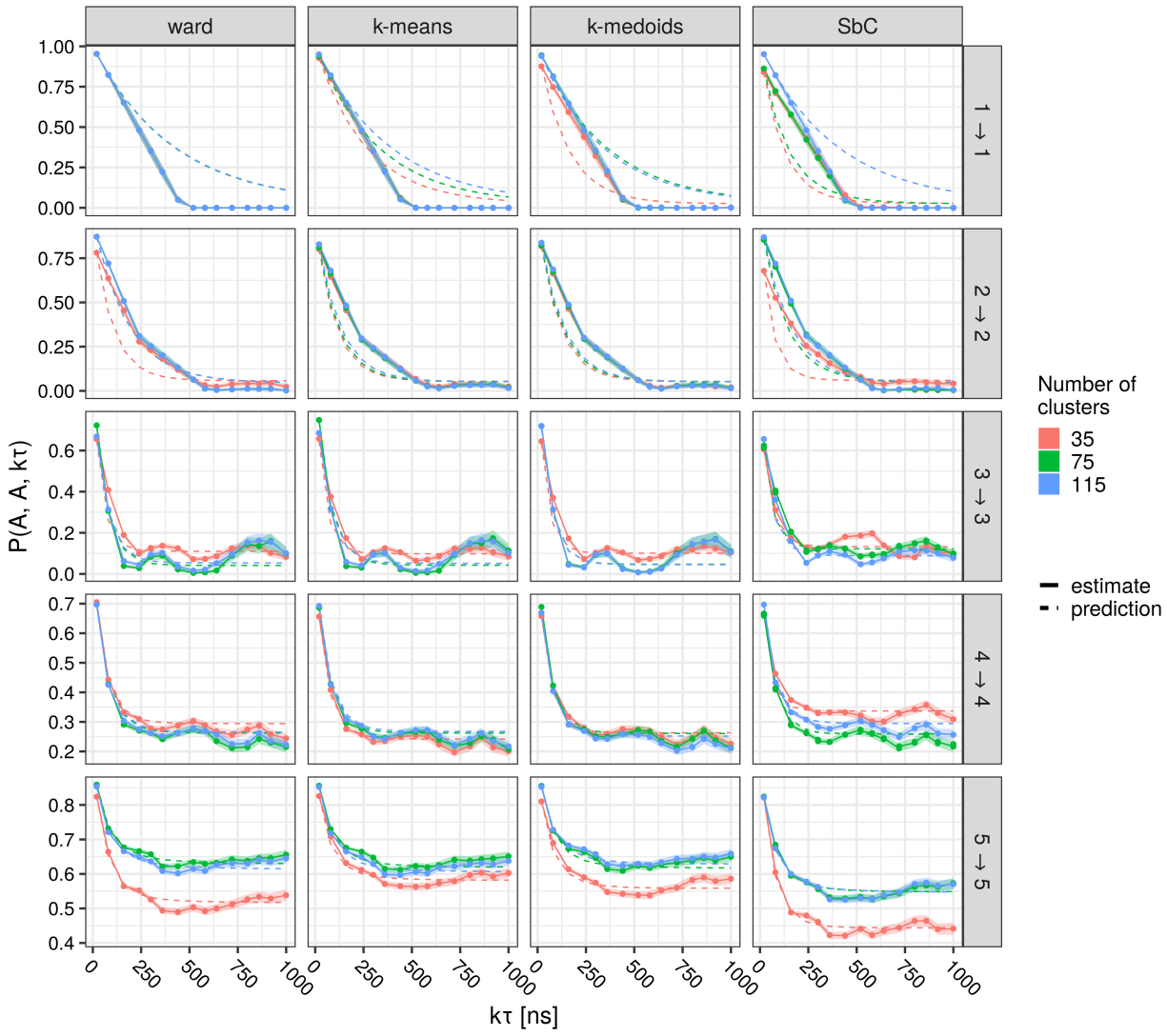

Figure S6. Chapman-Kolmogorov test of Beta3S for different clustering methods. Coarse-graining of the trajectories into five macrostates was performed with PCCA+.[2] Predictions are computed from an unconstrained maximum likelihood MSM with lag equal to 20 ns. Standard errors for the estimated values are shown as shaded regions.[3]

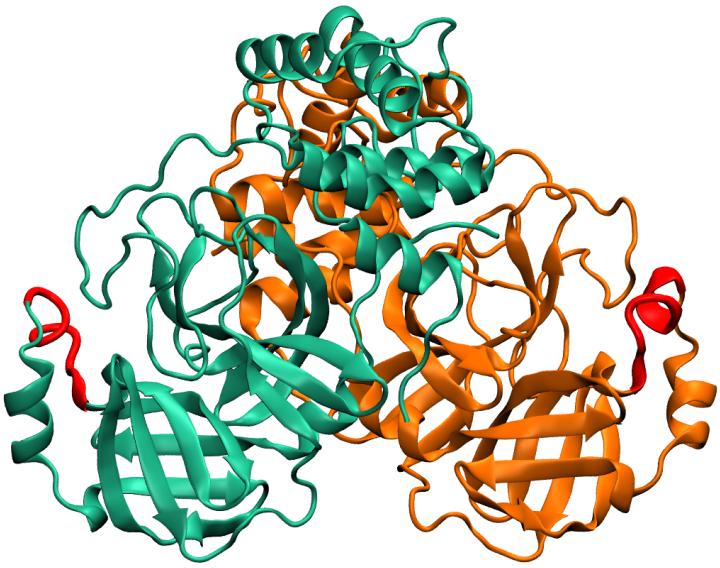

Figure S7. Cartoon illustration of the *apo* form of $3CL^{pro}$ (PDB entry 6Y84).[4] The color code for the chains is the same as in Fig. 7 in the main text. The residues highlighted in red are Cys44 to Asn51 (from bottom to top in the figure in both of the chains).

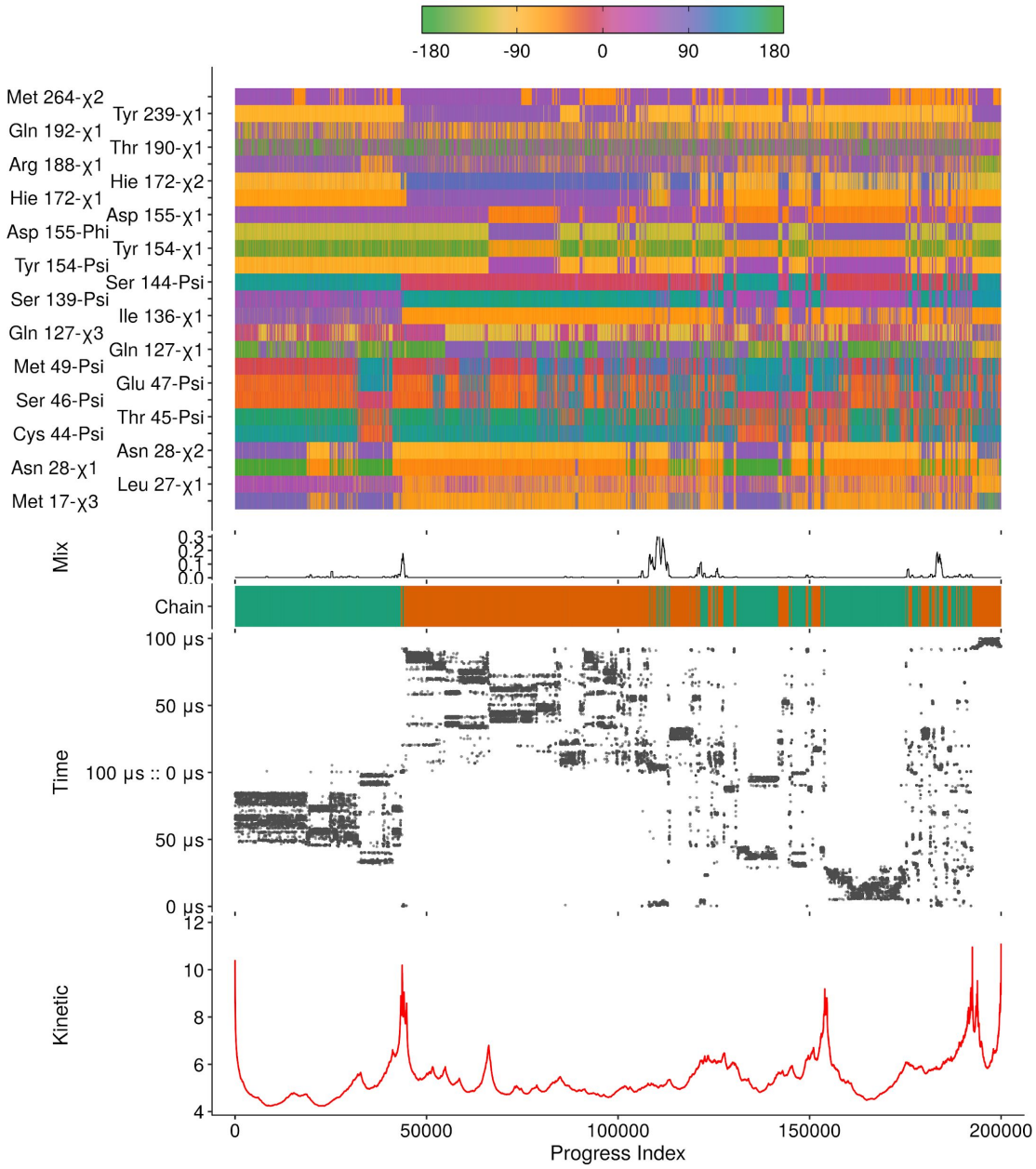

Figure S8. SAPPHIRE plot created from MD simulations[5] of the *apo* form of $3CL^{pro}$ (PDB entry 6Y84).[4] This figure is analogous to Fig. 7 in the main text except that here the 10 principal components encapsulating the largest amount of variance are used as features **in**cluding the first one. This first component contains high loadings for dihedral angles that separate the two chains (see 'Mix' annotation), *e.g.*, the χ-angles in residue Hie172.

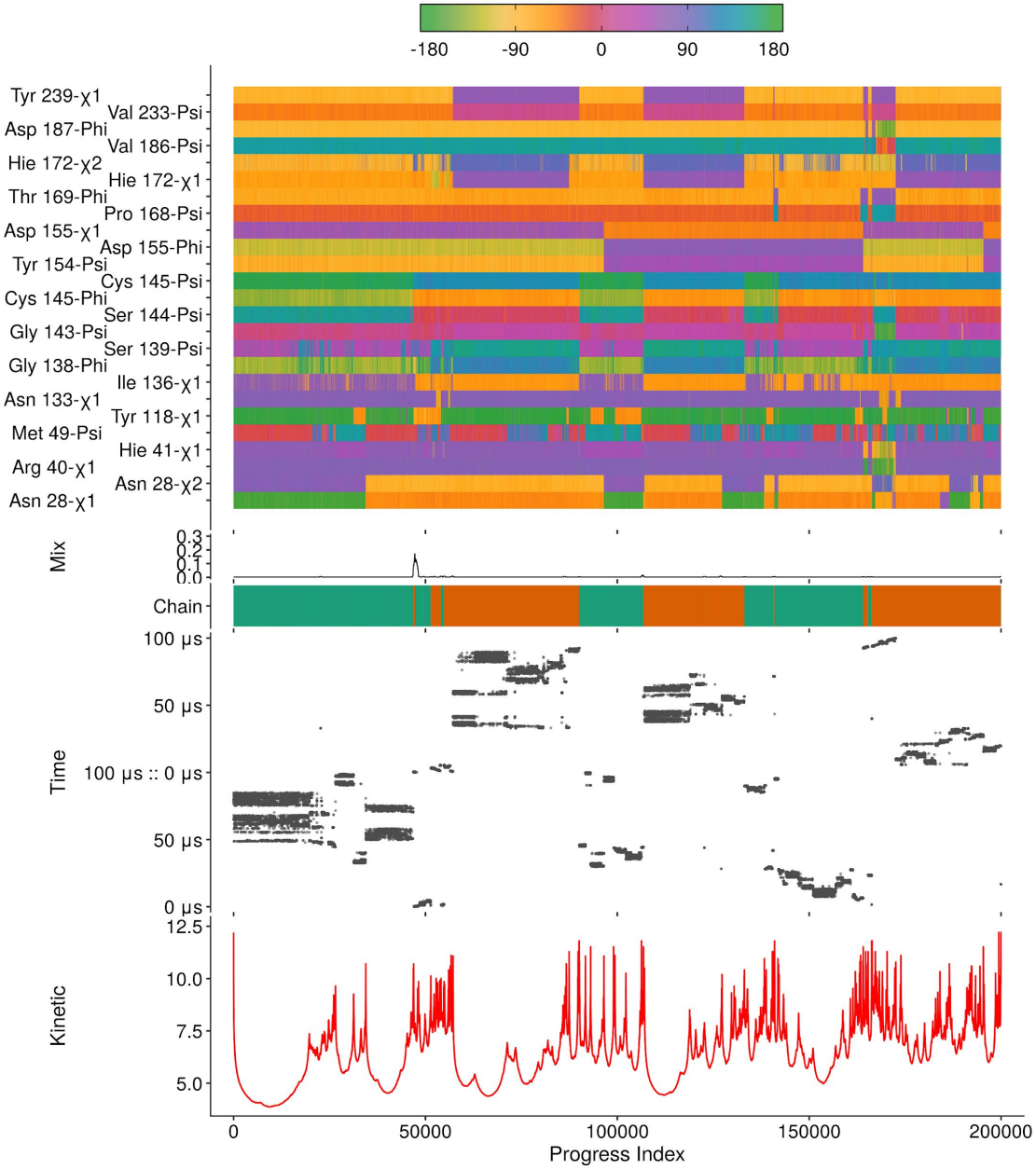

Figure S9. SAPPHIRE plot created from MD simulations[5] of the *apo* form of 3CL^{pro} (PDB entry 6Y84).[4] This figure is analogous to Fig. 7 in the main text except that simple time-lagged independent component analysis (tICA)[6] with an autocorrelation lag equal to 100 ns was used instead of principal component analysis. The source data were the same, *i.e.*, the data set of sine and cosine values of 865 dihedral angles, and the 10 components with the largest autocorrelation values were retained except that the first one was ignored (to be directly comparable to Fig. 7). The 25 dihedral angles shown are the ones that have the larges autocorrelation function value at the chosen lag. Because we concatenate the trajectories of the two chains in time, it is not necessarily surprising that tICA isolates static differences between the two chains as 'slow' modes, which are also visible in the dihedral angle annotations.

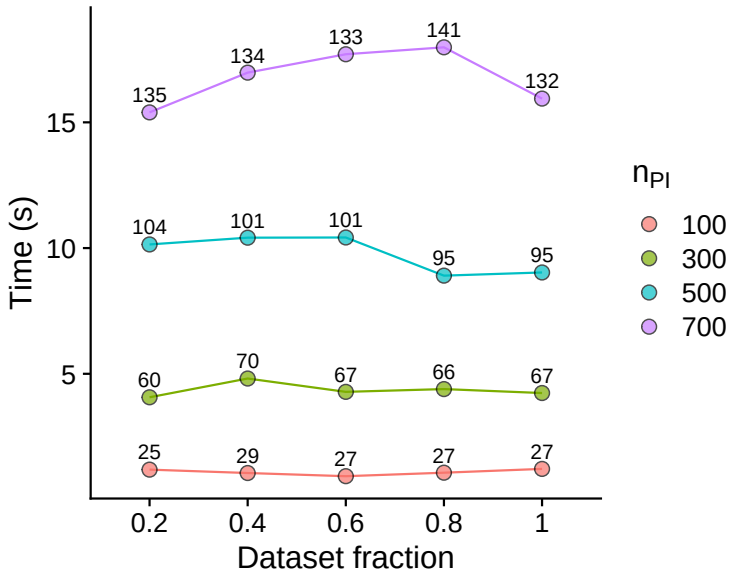

Figure S10. Computational cost of the SbC algorithm for different data set sizes and different values of $n_{PI}(= n_T)$ when applied to different subsets of the Beta3S trajectory (10^5 snapshots). The numbers of clusters identified by SbC are indicated. 'Time' on the y-axis refers to the elapsed CPU time of an R implementation of the SbC algorithm. The computational cost of the PI algorithm is not included. Both theory and data suggest that the cost of SbC does not depend substantially on data set size for the ranges tested. This implies that SbC is predicted to have a vanishing influence on the total computational cost as data set sizes increases. This is because the construction of the SAPPHIRE plot has a best-case scaling of $\mathcal{O}(N \log N)$ with data set size.[7]

References

(1) Vitalis, A.; Caflisch, A. Efficient Construction of Mesostate Networks from Molecular Dynamics Trajectories. *J. Chem. Theory Comput.* **2012**, *8*, 1108–1120.

(2) Röblitz, S.; Weber, M. Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification. *Adv. Data Anal. Classif.* **2013**, *7*, 147–179.

(3) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. Markov models of molecular kinetics: generation and validation. *J. Chem. Phys.* **2011**, *134*, 174105.

(4) Owen, C. D.; Lukacik, P.; Strain-Damerell, C. M.; Douangamath, A.,; Powell, A. J.; Fearon, D.; Brandao-Neto, J.; Crawshaw, A. D.; Aragao, D.; Williams, M.; Flaig, R.; Hall, D. R.; McAuley, K. E.; Mazzorana, M.; Stuart, D. I.; von Delft, F.; Walsh, M. A. SARS-CoV-2 main protease with unliganded active site (2019-nCoV, coronavirus disease 2019, COVID-19). 2020; `https://doi.org/10.2210/pdb6y84/pdb`.

(5) D. E. Shaw Research, Molecular Dynamics Simulations Related to SARS-CoV-2. D. E. Shaw Research Technical Data, 2020; `http://www.deshawresearch.com/resources_sarscov2.html`.

(6) Molgedey, L.; Schuster, H. G. Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Lett.* **1994**, *72*, 3634–3637.

(7) Blöchliger, N.; Vitalis, A.; Caflisch, A. A scalable algorithm to order and annotate continuous observations reveals the metastable states visited by dynamical systems. *Comput. Phys. Commun.* **2013**, *184*, 2446–2453.